1. 一种基于知识表示的生物医学实体链接方法，其特征在于，包括以下步骤：

步骤一、文本预处理

对于生物医学文本，首先提取出文本中所有待链接的生物医学实体提及，然后通过知识库查找实体提及对应的所有候选实体标识符 ID；仅保留排序前五的查找结果作为实体提及的候选 ID 集合；

步骤二、基于生物医学知识库的实体表示学习

（2.1）从知识库中抽取同一实体多种变体和不同实体同名的实体结构信息；

（2.2）将知识库中实体结构信息作为向量空间上的约束，采用自动编码机对实体提及表示和变体表示进行重构，从而学习实体 ID 表示；

自动编码机是基于两个约束：（i）实体 ID 表示是其各个变体表示的和，（ii）实体提及表示是其同名变体表示的和；定义实体提及表示为$m \in \mathbb{R}^d$，变体表示为$v \in \mathbb{R}^d$，实体 ID 表示为$s \in \mathbb{R}^d$；$v^{(i,j)}$是第 i 个实体提及$m^{(i)}$和第 j 个实体 ID$s^{(j)}$共有的变体；则自动编码机的两个约束可用如下公式表示：

$$s^{(j)} = \sum_i v^{(i,j)}$$

$$m^{(i)} = \sum_j v^{(i,j)}$$

自动编码机由两部分组成，即编码器和解码器；编码时，编码器按照实体提及→变体→实体 ID 的顺序进行；其中，实体提及表示$m^{(i)}$初始化为其组成单词的预训练词表示的平均值，变体表示$v^{(i,j)}$通过引入一个对角矩阵$E^{(i,j)} \in \mathbb{R}^{d \times d}$对实体提及表示$m^{(i)}$进行分解获得；然后，由对应变体表示的加和获得实体 ID 表示$s^{(j)}$；编码过程的公式如下：

$$s^{(j)} = \sum_i v^{(i,j)} = \sum_i E^{(i,j)} m^{(i)}$$

$E^{(i,j)}$是一个对角矩阵，满足条件$\sum_j E^{(i,j)} = I_n$，其中I_n是一个单位矩阵；

解码时，解码器按照实体 ID→变体→实体提及的顺序进行；通过引入另一

个对角矩阵$D^{(j,i)} \in \mathbb{R}^{d\times d}$将编码获得的实体 ID 表示$s^{(j)}$分解为各个变体表示$\bar{v}^{(i,j)}$，然后再由同名变体表示的加和重构实体提及表示$\bar{m}^{(i)}$；解码过程的公式如下：

$$\bar{m}^{(i)} = \sum_j \bar{v}^{(i,j)} = \sum_j D^{(j,i)} s^{(j)}$$

对角矩阵$D^{(j,i)} \in \mathbb{R}^{d\times d}$同样满足条件$\sum_i D^{(j,i)} = I_n$，其中$I_n$是一个单位矩阵；

（2.3）定义一个重构误差函数来训练自动编码机的参数，其公式为：

$$Loss = \alpha \cdot \| \sum_j (D^{(j,i)} \sum_i E^{(i,j)} m^{(i)}) - m^{(i)} \| + \beta \cdot \| E^{(i,j)} m^{(i)} - D^{(j,i)} s^{(j)} \| \quad \forall i, j$$

该重构误差函数由两部分组成，一个是要求解码出的实体提及表示$\bar{m}^{(i)}$与输入的提及表示$m^{(i)}$对齐，即$\sum_j (D^{(j,i)} \sum_i E^{(i,j)} m^{(i)}) \approx m^{(i)}$；另一个是要求解码器得到的变体表示$\bar{v}^{(i,j)}$与编码器得到的变体表示$v^{(i,j)}$对齐，即$E^{(i,j)} m^{(i)} \approx D^{(j,i)} s^{(j)}$；

通过最小化该重构误差函数，使得实体结构信息被嵌入到实体 ID 中，得到学习后的实体 ID 表示；α, β为权重系数，且满足$\alpha + \beta = 1$，用于控制两部分对齐的平衡；

步骤三、基于知识表示的生物医学实体链接

利用步骤二学习获得的实体 ID 表示，对步骤一抽取出的生物医学实体提及进行消歧，获得在特定上下文中实体提及对应的唯一 ID；构建基于知识表示的实体消歧模型，该模型通过注意力机制和门机制融合文本语义表示和实体 ID 表示，从而预测实体提及被链接到当前候选实体 ID 的概率；具体过程如下：

（3.1）通过嵌入层，将待链接实体提及的候选实体 ID，和它的左侧上下文、右侧上下文分别映射到向量空间，获得候选 ID 表示s和左、右侧上下文词向量序列$C^L \in \mathbb{R}^{d\times n_2}$和$C^R \in \mathbb{R}^{d\times n_2}$；

（3.2）C^L和C^R分别输入给一个门递归单元神经网络 GRU，获得在第 t 个时间步输出的隐层表示h_t^L和h_t^R如下：

$$h_t^L = GRU^L\left(C_t^L, h_{t-1}^L\right)$$

$$h_t^R = GRU^R\left(C_t^R, h_{t-1}^R\right)$$

对于一段词序列的语义信息，其中每个词相对于候选ID的重要性是不同的；所以提出基于知识表示的注意力机制，利用候选ID表示计算每个时间步隐层表示的归一化权重α_t，计算公式如下：

$$\alpha_t = \frac{\exp\left(e_t\right)}{\sum_{i=1}^{t}\exp\left(e_i\right)}$$

$$e_t = \tanh\left(W_a \cdot h_t + V_a \cdot s + b_a\right)$$

其中，h_t是h_t^L或h_t^R；$W_a \in \mathbb{R}^{1\times d}$，$V_a \in \mathbb{R}^{1\times d}$和$b_a \in \mathbb{R}^{1\times 1}$均是模型的参数，在训练过程中进行调优；tanh为双曲正切激活函数；通过一个前馈神经网络建模候选ID表示s与上下文各个时间步的隐层表示h_t，获得二者的关联得分e_t；之后，利用softmax函数对得分e_t进行归一化得到隐层表示的权重α_t；

接下来，对GRU隐层表示的整个序列作加权求和操作，使结构信息编码的候选ID表示与上下文语义表示融合，公式如下：

$$o = \sum_t \alpha_t h_t$$

其中，o表示左上下文表示o^L或右上下文表示o^R；

（3.3）注意力机制分别应用于左、右侧隐层表示从而获得的左、右上下文表示，并通过一个门机制来进行动态控制，让实体提及的最终上下文表示z获得充分的学习，计算公式如下：

$$z = g \odot o^L + \left(1-g\right)\odot o^R$$

$$g = \sigma\left(W_g \cdot o^L + V_g \cdot o^R + b_g\right)$$

其中，W_g, V_g, b_g是待训练的参数；$\odot$表示逐元素相乘；g是权重，通过将左右上下文o^L和o^R输入一层全连接层并通过sigmoid激活函数σ获得的；

（3.4）将实体提及的上下文表示z和候选ID表示s拼接输入给分类器；分

类器由具有 ReLU 激活的两层全连接神经网络 FC 和一个 sigmoid 输出层组成，公式如下：

$$pr_1 = relu\left(W_1 \cdot \left[z;s\right] + b_1\right)$$

$$pr_2 = relu\left(W_2 \cdot pr_1 + b_2\right)$$

$$p = \mathrm{softmax}\left(W_3 \cdot pr_2 + b_3\right)$$

其中，$W_1, b_1, W_2, b_2, W_3, b_3$是待训练的参数，$[;]$表示拼接操作，$p$ 为实体提及被链接到当前候选 ID 的概率；

（3.5）基于知识表示的实体消歧模型通过一个二元交叉熵损失函数进行训练，公式如下：

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log p_i + \left(1 - y_i\right)\log\left(1 - p_i\right)\right) + \lambda \| W \|$$

其中，n 是训练样例个数，y_i 是第 i 个样例对应的正确标签，p_i 是第 i 个样例的预测概率，$\lambda \| W \|$为训练参数的正则项；

生物医学实体消歧模型为每个实体提及与其候选 ID 进行打分并排序，选择得分最高的候选 ID 作为最终的链接结果。